**Lead Scoring Case Study**

**Summary:**

We started with importing the required libraries like pandas, numpy, matplotlib, etc. After this we imported the data set in a variable called leads using pandas and checked its different attributes like shape, info, describe, etc.

After doing all the above steps we started off with the data prepossessing steps like handling missing values, checking outlies, standardization, scaling, checking imbalance data. First, we checked the percentage of missing value in all the column and dropped all the columns which had more that 35% of the total missing values as if we try to impute this values the model may give us biased results which may not help the business problem. By removing these columns we have left with 31 columns. Then we started to handle the missing values in categorical column. Here we imputed the NAN values with the mode of the respective categorical column. After this we checked for the numerical column for the missing values but we didn’t have much numerical column which had missing values only ‘TotalVisits’ column had around 1.5% of the missing values hence we dropped those rows as it will not impact the model much.

After missing value handling we checked for the overall conversion rate which came out to be 38.5%. After this we checked in each of the remaining column for the imbalance or skewed data. By doing this we found out that there were 15 columns who had imbalance data hence we removed this column along with some sales related columns like ‘Last Activity', ‘Lead profile’, ‘Lead Number', etc. After removing these columns I we were left with 10 columns.

After this we checked for outlies in numerical column and found out 2 columns had outlies in it ‘TotalVisits’ and ‘Page Views Per Visit’. We handled the outliers by removing the values over Q3 for upper limit and below Q1 for lower limit.

We did some standardization in the few columns while performing EDA and founds out some basic insight like:

* We can see that overall convertion rate of the whole data is 38.5%
* Maximum number of leads are generated by Google and Direct traffic,
* Conversion Rate of reference leads and leads through welingak website is high
* As we can see the maximum lead is generated from Unemployed Category.
* The conversion is high in case of Working professional.
* Leads spending more time on the website are more likely to be converted.

After all these Steps we converted all the categorical columns into dummy variables and dropped all not require columns from the data which left us with 34 columns.

After this we started preparing the modal. For this we did train-test split. We took 70% of the data to train the modal and 30% of data to test the model. After splitting we did scaling numerical columns which had wide ranges of value and by using StandardScaler we scaled them down between -1 to 1. This was a very important step as if we would have left it like that the modal would have become partial toward this wide range of values.

After scaling the columns next step we took was to automate the features selection process and for this we used RFE. RFE helped use to automatically select the most significant columns for our modal and by using RFE we selected 15 features. Then by using stats modal library we created our modal. After creating the model the next we checked p-vales and VIF of all the 15 variables which explained us the significance and multicollinearity of each features. We dropped the variable having more than 5% of P-value and VIF of more than 5. This was an iterative process till the time we had achieved all the significate and non-multicollinear features.

After this we started predicting the probabilities using the modal for training dataset and assigned a lead score to all the rows from 0 to 100 as 0 means the least likely to convert and 1 means most likely to convert. After this we selected an arbitrary cutoff point as 0.5 which meant that if the predicted probability is less than 0.5 then that customer will not convert and if the if the predicted probability is more than 0.5 then the customer will convert.

Then we found confusion matrix and calculated accuracy, sensitivity, specificity, precision and recall for all traing set. But the values we got was not reliable as the arbitrary cutoff point we took was at random and we need a more optimal cutoff. Hence to find the optimal cutoff we calculated the accuracy, sensitivity and specificity for different cutoff like 0.0, 0.1, 0.2 so on till 0.9 and plotted them and took the point where curve of all the metrics intersected. The optimal cutoff point came out to be 0.3.

Then again we found confusion matrix and calculated accuracy, sensitivity, specificity, precision and recall for train set for the cutoff point 0.3 and we found the below values.

|  |  |  |
| --- | --- | --- |
| Predicted  Actual | Not\_Conv | Conv |
| Not\_Conv | 3016 | 866 |
| Conv | 570 | 1815 |

* As we can see from the above matrix true positive according to modal is 1815.
* True negative according to modal is 3016.
* False positives according to modal is 570.
* False negatives according to modal is 866.
* Accuracy= 0.77, Sensitivity= 0.761, Specificity= 0.776
* False Positive Rate= 0.223, Positive predictive value= 0.676, Negative predictive 0.841
* Precision= 0.767, Recall= 0.761

We also plotted precision and recall tradeoff to assess the modal for training data set.

After this we started to assess the modal on test dataset. For this we first did the scaling of test data and then with our modal we predicted the probabilities of getting converted and also assigned the lead score to each of the customers from 0 to 100.

Then we took the same optimal cutoff 0.3 to converge a customer to 0 or 1. we found confusion matrix and calculated accuracy, sensitivity, specificity, precision and recall for test set and we found the below values.

|  |  |  |
| --- | --- | --- |
| Predicted  Actual | 0 | 1 |
| 0 | 1304 | 372 |
| 1 | 227 | 783 |

* As we can see from the above matrix true positive according to modal is 783.
* True negative according to modal is 1304.
* False positives according to modal is 227.
* False negatives according to modal is 372.
* Accuracy= 0.776, Sensitivity= 0.775, Specificity= 0.778
* Precision= 0.677, Recall= 0.775

So we concluded the below points after the modal is prepared.

* The Model seems to predict the Conversion Rate very well and we can make calls based on this model.

Train Data:

* Accuracy : 78.55%
* Sensitivity : 62.64%
* Specificity : 88.33%

Test Data:

* Accuracy : 77.69%
* Sensitivity : 77.52%
* Specificity : 77.80%
* Team should approach to all those costumes how are spending more time on the website.
* Team should approach to all those costumes how are working professionals.
* Team should approach to all those costumes how are filling up the add forms.
* Team should approach to all those costumes how are coming from Welingak website.